**Working title**

Third generation sequencing reveals conserved restriction-modification target motifs in wild isolates of *Borrelia burgdorferi* *sensu stricto* and target site diversity across *Borrelia* species.

**Abstract**

Text

**Introduction**

Text

**Results**

***I. Nanopore sequencing***

***II. Identification of modified motifs from third generation sequencing data***

Text

***III. Genome-wide distribution of modified bases***

Text

***IV. Restriction-modification systems across the Borrelia genus***

Text

***V. Phylogenetic Analysis of R-M genes***

Text

**Discussion**

Text

**Methods**

***I. Sample collection and culturing***

*Ixodes scapularis* ticks were collected from Rockefeller State Park Preserve in Pleasantville, NY, and Caumsett State Historic Park Preserve in Lloyd Harbor, NY in October of 2022 by dragging canvas sheets along the ground in forested areas. Ticks were dissected and contents of the midgut were cultured in BSK-H complete medium (Sigma) with 6% rabbit serum and incubated at 34 C. Cultures were checked for the presence of spirochetes after 2 weeks using phase contrast microscopy. The density of positive cultures were calculated using a Petroff-Hausser counting chamber, followed by serial dilutions to get single strain isolates prior to PCR and genotyping.

***II. PCR and Genotyping***

PCR was performed using a pair of primers designed for targeting the full length ospC gene. The forward primer, 5’-AATAAAAAGGAGGCACAAATTAATG-3’ and reverse primer, 5’-ATATTGACTTTATTTTTCGAGTTAC-3’ target the intergenic spacer regions in the genes flanking ospC. PCR reaction mixtures of 20 ul per sample were prepared as follows: 16 ul ddH2O, 2 ul 10X Roche FastStart buffer (Roche Diagnostics, California, USA), 0.4 ul of 50X deoxynucleoside triphosphate (dNTP), 02. Ul of 100X Roche FastStart Taq DNA polymerase (Roche Diagnostics, Mannheim, Germany), 0.4 ul of 50X primers, and 1 ul of sample. Reaction mixtures were heated at 95C for 3 minutes, amplified for 36 cycles at 95C for 15 seconds, 60C for 15 seconds, 72C for 1 minute, and incubated at 72C for 3 minutes. Samples were kept at 4C after finished until collected. Gel electrophoresis was performed on the PCR products on a 1.5% agarose gels stained with ethidium bromide and viewed under UV light. Bands corresponding to the molecular weight of ospC were cut out of the gel and the DNA was purified using GeneJet Gel Extraction Kit (Thermo Scientific) following the manufacturers protocol. 8 ul of the purified PCR product (with concentrations between 10-20 ng/ul) were sent out for Sanger sequencing (Eton Bioscience, Inc, California, USA).

Blastn was used to search the sequenced PCR products against a database of 64 published ospC sequences (21 sensu stricto sequences, 43 sensu lato sequences). Samples with an e-value of 0 for a particular ospC serotype were considered as single ospC strain types.

***III. DNA isoloation***

[need to get protocol from Alvarado lab]

***IV. Sequencing and data pre-processing***

Oxford nanopore (ONT) sequencing was performed on a GridION sequencer using FLO-MIN114 flow cells. DNA libraries were prepared using the ligation sequencing kit V14 (SQK-LSK114) following the manufacturers protocol. A minimum Q score threshold of 9 was used to filter out low quality reads, and the top 1% of the longest reads were discarded as outliers. ONT’s Dorado software version 0.5.1 was used for base calling. The dna\_r10.4.1\_e8.2\_400bps\_hac@v4.3.0 model was used for canonical base calling simultaneously with modified base calling using the dna\_r10.4.1\_e8.2\_400bps\_hac@v4.3.0\_6mA@v2 model. Reads were mapped to reference genomes using Minimap2. The resulting bam files were sorted and indexed using Samtools, and modified base calls were summarized using the Pileup function in ONT’s Modkit software.

***V. Identification of modified DNA target motifs***

Reads were aggregated for each adenine position in the genome with a minimum of 10X coverage, and counts were tallied using the modified base calling results to determine the methylation status of each adenine in the genome. 11 bp sequences for positions with modified calls on greater than 50% of the reads were extracted (5 bp up and downstream of modified adenine). To detect enriched patterns within the set of 11-mers, frequencies for all possible k-mers of length 4 to 10 (containing the modified adenine) were calculated for the set of sequences. Candidate target sequences were then constructed from sets of similar patterns occurring at high frequencies by assigning ambiguous nucleotides to the variable positions. These consensus sequences were validated by determining the percentage of adenines that are modified within areas of the genome that match the specific pattern. Restriction modification targets are expected to be methylated at very high frequencies (>85%).

***VI. Genome wide methylation analysis***

Custom scripts were used to assess global methylation patterns in each sequenced sample. Methylated positions were annotated by aligning each position to genomic features in the published Genbank gbff files. Sliding window analyses were performed across every amplicon to detect methylated hot spots and regions void of DNA methylation. Potential enrichment or avoidance of the inferred R-M target motifs within coding sequences was assessed by comparing the observed motif counts to counts in synonymously shuffled sequences (n=100), and performing one sample t-tests for each motif and each coding sequence.

***VII. Restriction modification system search***

For a comprehensive search on restriction modification systems harbored within the genomes of *Borrelia sp.*, a complete set of coding sequences from each of 90 *Borrelia* genomes (see table ) were searched against the REBASE Gold Standard Protein sequences (citation), a set consisting of 3,588 experimentally validated restriction modification system genes, using Blastx (citation?). An e-value threshold of e-30 was used in the search, and homologs to the proteins in the REBASE set were identified as the hit with the lowest e-value that covered at least 75% of the subject sequence.

***VIII. Multiple sequence alignments and phylogenetic tree construction***

Coding sequences for the putative R-M system components matching to the same REBASE homologs were translated into peptide sequences using Bioseq (citation). Alignments of the peptide sequences were generated using MUSCLE (citation), and Bioalign (citation) was used to compute sequence similarity and percent identity. FastTree (citation), which applies the approximately-maximum-likelihood method, was used to infer phylogenetic trees from the peptide alignments. The R package ggtree (citation) was used to visualize and annotate the trees.

**Data and Code Availability**

The in house tool package made for identifying R-M target motifs and characterizing the methylomes is available on github (link to repo). \*Nanopore sequence data\*

**Acknowledgements**

Text

**References**

Text

**Tables**

**Table 1: Genomes used in R-M search and results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Species** | **Strain** | **GenBank ID** | **RM Homologs** | **RMS type** | **Genomic Location** |
| B. burgdorferi | B31 |  | CjeNIII, CchII | IIG | lp25, lp56 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| B. afzelii |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| B. garinii |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**Table 2: RM search summary by *Borrelia* group**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Homolog | RM Type | Modification | Target | Presence | | | | |
| **LD** | **RF** | | **REP** | |
| CjeNIII | IIG | 6mA | GKAAYG | ✓ | |  | | ✓ |
| CchII | IIG | 6mA | GGARGA | ✓ | | ✓ | | ✓ |
| Hhal | MTase | 5mC | GCGC | ✓ | |  | |  |
| GSP928I | MTase | 6mA | GATC |  | | ✓ | |  |

**Figures**

**Figure 1: RM System Search workflow**

A diagram of a protein complex

Description automatically generated with medium confidence

Figure 2: Type IIG Restriction Modification Gene Diagram

**Figure 3:** CjeNII homolog gene tree

A circular chart with different colored circles

Description automatically generated with medium confidence

**Figure 4:** CjeNII *Bb s.s.,* homolog gene tree annotated by snp group

A diagram of a number of dna

Description automatically generated

**Figure 5:** CchII homolog gene tree

A close-up of a graph

Description automatically generated

**Figure 6:** Nanopore sequencing workflow

A screenshot of a computer

Description automatically generated